Date,Time & Location:	**Architecture Workspace and V/CDE Workspace Liaison Meeting** 5/18/04, 1-3 PM, NCICB, 6116 Executive Blvd. Rockville, MD (by teleconference)
Attendees:	Roll call was performed with the following participants attending: **Architecture Liaisons** Fred Hutchinson • Robert Robbins Ohio State • Scott Oster • Tahsin Kurc **V/CDE Liaisons** Mayo • Harold Solbrig UC-Davis • Cecil Lynch Albert Einstein • Xin Zheng NCICB • Frank Hartel **Other Attendees** Washington University • Rakesh Nagarajan University of Pittsburgh • Jim Harrison University of Hawaii • Leo Cheung Jackson Lab • Jim Kadin City of Hope • Joyce Niland • Hemant Shah • Jennifer Neat OHSU • Lara Fournier EMMES Corporation • Claudia Valmonte • Ryan Campbell

NCICB

- Peter Covitz
- Leslie Derr
- John Qu
- Juergen Lorenz

NCI/OC

- Larry Wright
- Margaret Haber

Fred Hutchinson

- Dan Geraghty

Coldspring Harbor

- Michael Townsend

University of Wisconsin

- Rhoda Arzoomanian

SAIC

- Kathleen Gundry

BAH

- Arumani Manisundaram
- Christine Richardson
- Mike Keller

Agenda Item #1:

Goal of Meeting/Introduction

Peter Covitz opened the meeting by stating that the Architecture/VCDE group needs to define caBIG 'compatibility', and what it means for caBIG Workspace products or artifacts to be caBIG-compliant. This does not mean drafting scenarios, but providing guidance across the caBIG community.

- The two Cross-Cutting Workspaces need to be working together at all times.
- The Architecture WS has decided to break into sub-groups, one of which is Information Architecture. They are involved with determining how data are portrayed in the space as well as re-distributing the meaning of those data in a grid-type fashion. This corresponds with the issues surrounding domain models and data representation.
- It is hoped that we can put some basic recommendations together for developers relatively quickly depending on the outcome of this meeting. We will spend a fair amount of time on Agenda Item #2.

- A broad caBIG requirement might be characterized in the following way: 'When I sit down to my caBIG console I would like to be able to the following.'
    - Determine what data are available
    - Determine what the data mean
    - See how they are represented
    - Determine how the data fit into the broader space of biomedical information
- The goal of this liaison group meeting is to determine how the answers to these questions are going to be formalized.

Using the example of caCORE as a departure:

- At model level, we use UML (modeling language-formal way of describing entities or classes). UML is accessible; additionally it is a formalism that can be fed into other software tools.
- UML is great for describing a broad domain of interest. But we need, also, to get down to nuts and bolts; UML does not allow us to go another level of granularity. CDEs are little pieces represented in UML. The next level down in granularity are the terms (or values) that populate the data with (controlled vocabulary or ontology)
- Essentially, caCORE has broken these down into three chunks with different levels of granularity:
    - UML
    - CDE
    - Terms to define data and semantic standards (Vocabs)

Agenda Item #2:

Metadata and Domain Models

Harold Solbrig introduced the SAGE Project at Mayo. This project is trying to create interoperable guidelines to be able to reference data in a neutral fashion.

- The terminology was the key set of definitions throughout the model. The terminology was key to coming up with attributes and possible values for the attributes. (Using Protégé.)

- Found it was necessary to anchor the terminology all the way throughout the spectrum. (SNOMED CT)
- Vocabularies and common data elements need to be the definitive glue for organizing definitions. An organized set of definitions needs to permeate the information model (e.g. a Stage IV tumor needs to mean the same thing in every case).

Bob Robbins briefly gave his background emphasizing that he speaks from a basic science frame of reference and not a clinical one. His experience tells him that if caBIG is to succeed then the following issues need to be addressed:

- 'Meaning' changes over time. Scientists can agree on the term, but multiple meanings or definitions evolve.
- Requirement for some sort of universal naming authority.
- 'Interdatabase referential integrity'. Need to build infrastructure like a cascade or a notification system that tells the 'target' database about foreign keys.

Cecil Lynch next brought up the topic of semantic drift.

- Cecil Lynch: A term that has experienced extensive semantic drift needs to be handled as a different (or new) term.
- Cecil Lynch: Doesn't see how multiple definitions would work out.
- Peter Covitz: caBIG will have a forum to reconcile these issues.
- Semantic drift, proposed multiple definitions, way to handle it if there is that much semantic drift then it has to be a different definition.
- Cecil Lynch: One example is with the words gene and locus. These are basic concepts that have acquired slightly different meanings to different people over time.
- Harold Solbrig: Ontologies and terminologies are never nicely partitioned, we need to manage some fuzziness; what we want to do as much as possible is agree with what the thing is.
- Peter Covitz: Permeation of the information system with vocabulary is key. Five years pass, you retrieve the operational definition that was put in the system when the data were collected and then you determine if the definition is still valid.

- Harold Solbrig: When you give a definition to a data element, it is quite possible that the thing might be identified in many ways. Need to publish a standard naming mechanism. That's where you need a shared information model.

Cecil Lynch returned to the conversation to the topic of referential integrity.

- Cecil Lynch: How can we maintain concurrency across disparate databases with different key structures? A solution to this would be to look to the metadata registry.
- Harold Solbrig: At bare minimum, it is crucial to come up with common names. Next step is to be able to publish or make available info that others need or find useful.
- Rob Robbins: Need to regard 'referential integrity' foreign key vs. primary key. Need to provide unambiguous mapping, and when necessary, keep everyone notified of changes.
- Peter Covitz: This problem of naming is everywhere on the web (Universal Reference identifiers is something we could potentially use).
- Rob Robbins: Another problem is that scientific data objects degrade. Most commonly, over time, things are either lumped together or split. For example, one gene may become actually two genes with an intervening region, or a common cause is determined for two diseases and they are subsequently lumped together as one disease entity. How shall we deal with this?
- Frank Hartel: These problems have been tackled (by the NCICB team work). Using SNOMED CT and the NCI Thesaurus pathway they maintain a recorded history of the entire life cycle of every concept (or term) in vocabulary (with predecessor terms). They can determine what the current terminology is and what all the predecessors were, by date. These are available both at concept and term level with SNOMED, and can be leveraged.

ACTION ITEM: Build a deliverable that is a series of 'guiding principles' for what constitutes a work product that is caBIG-compliant. Begin with caCORE and determine if it is sufficient and go from there.

- Joyce Niland: Is mapping of CDEs to SNOMED and LOINC a guiding principle? Should new terminology be created only if the CDEs do not exist?

- Peter Covitz: In general, are people comfortable with the caCORE CDE Browser database?
- Peter Covitz: (EDRN, from the Hutch-FHRCR, is used as an example for discussion.) EDRN system is CDE-driven for semantic continuity. There is a software infrastructure on top; distributes queries across sites. It is not 'open', however. The EDRN represents a good case study for interoperability. It does not provide any real formal modeling environment to establish relationships between data elements. That's where UML comes in, to allow for higher order relationships across data elements. Is the CDE sufficient to define interoperability? Do we need other models, such as classes? For example, not just a list of genes, but genes in a certain pathway?
- Margaret Haber: Semantics need to permeate at all levels. We should only create CDEs that fill gaps, then, create the unifying semantic glue.
- Peter Covitz: Do we need UML-like relationships?
- Harold Solbrig: Is an advocate of formalization, however, it is not totally obvious how formal we have to go. The advantage of UML, is that when you put things together graphically, it helps to clarify some bad misunderstanding which can occur. Would advocate for representing model-like relationships across data elements.
- Cecil Lynch: Agrees completely. Context plays a significant role. We should centralize controlled vocabularies and integrate terms, not create new ones. For example, NCI has anatomical terms; there are terms in SNOMED that are not in NCI, so add the terms from SNOMED to NCI rather than 'creating' them.
- Harold Solbrig: There is not a clear boundary between the information model and the vocabulary mode. There are things that need to be clarified.
- Peter Covitz: Do we need a universal identifier system that goes all the way down to the term or data object? Or is it sufficient instead to have a universal identifier system to define a class? We can define what a gene is, but not specific genes. Is that a tolerable amount of specification or do we need more?
- Rob Robbins: We want decent reliability in quality of data objects in caBIG. We need to understand the distinction

between using the grid for purposes of information retrieval vs. computational analysis.

- Peter Covitz: How should we characterize a caBIG-wide universal naming (primary identifiers) system?
- Rob Robbins: In thinking about a way to come up with unique identifiers for caBIG sources, we should make sure that the source and the object are unique. Most universal ID systems are multi-part (e.g. zipcodes, ISBN).
- Harold Solbrig: CaBIG can have control over some data classes, but others are outside external control. With certain information models, e.g. gene identifiers, you are at the mercy of what is the accepted practice. Need to cope with multiple identification schemes.
- Jim Kadin: You have to have identifiers of the individual genes; the only way to refer to the gene is by an identifier. Gene, sequences, clones, and snips have to have identifiers. Multi-part identifiers should be used once you start applying attributes. You don't want to change primary identifier if the object type changed. Adopt identifiers that already exist.
- Frank Hartel: We adopt naming conventions already assigned from outside sources, e.g. genes are consistently named by outside authorities. When UNIPROT comes along we will look to that for identifiers.
- Cecil Lynch: The ID of an object should be kept distinct from an object attribute or association.
- Peter Covitz: There appears to be an acceptance of the idea that using established object IDs is OK.
- Rob Robbins: Not completely. Borrowing a primary key from other external authors is not acceptable. But to use them as a reference is acceptable. The external ID can prepend or add to the caBIG prefix.
- Cecil Lynch: Need to understand how to handle overlap? This is a very difficult issue.
- Peter Covitz: How far do we want to go on rules for Object Permanence? There may be classes of data where immutability of naming is important.
- Peter Covitz: Want to return to the topic of vocabulary mapping through the information model. If we have classes, and attributes of those classes, and instances of those classes (which are the data objects), what metadata do we need to

provide a caBIG compatible grid service? Every data class, as well as every attribute, has to have a definition. Some of these attributes will include: discrete vs. continuous, length (restrictions), alpha vs. numeric, etc. What of these attributes fall outside of our needs, i.e. are not semantic?

- Harold Solbrig: We may not necessarily have a well-specified definition, but can reference an appropriate definition system.
- Peter Covitz: What level do you need to define the values themselves? Do we demand this? With the CDE tool there is a field to define your valid values. In caDSR world, for example, M and 1 could both mean male; F and 2 could mean female. Both value systems have unique meaning…even though they are different.
- Margaret Haber: LOINC mapping is in the NCI Metathesaurus. The UMLS includes LOINC so the base code is in UMLS.
- Harold Solbrig: Is it your intention to assign an NCI code to every LOINC code?
- Margaret Haber: No. We end up with NCI code when things are not represented in tUML code.
- Harold Solbrig: When you use Sage/HL7, 90% of what you need is there.
- Margaret Haber: We can refine the mappings when you load the local terms.

Peter Covitz provided a summary of the meeting's discussion.

- The vocabulary environment and semantics need to permeate the information model.
- We need a strategy for building concept history and to guard against semantic drift.
- We need to have a universal identifier system (to combine source and object identifier, and it probably needs to be consistent with HL7)
- We need a strategy to manage concept and other changes.
- Definitions must accompany data.
- We need standardized naming conventions for data classes.
- We need some type of higher modeling representation (UML or something comparable) that allows one to graphically link data classes and define relationships.
- Multiple definitions of same term need to be managed.
- We need a mechanism to define translational services from

	local to shared, universal definitions.
Agenda Item #2:	**<u>Management of Vocabulary and CDE sources</u>** The following questions were raised by Peter Covitz. • Do we create UML repository for cancer centers to deposit their models? • Do we use caDSR? • Do we use EVS? • Do we charge Mayo with development of the next generation vocabularies server? • Do we envision a federation of these services? • What is practical for caBIG to deploy? • Xin Zheng: There could be a hierarchy with central authority. Global services managed by global authority, then build in sub-group authorities (branches), with a distributed approach at that level. • Harold Solbrig: Would be more inclined to decide these things based on review of use cases. • Peter Covitz: There needs to be a central system to generate identifiers (UID). • Harold Solbrig: One approach to building an ID system is delegation. • Peter Covitz: Deployment topology is still a little early; we need to know more. • Harold Solbrig: About both use cases and data volume. • Peter Covitz: Arumani and Christine will compile meeting notes and distribute them. We will have informal discussion with liaisons to determine the action items and deliverables, such as white papers, and high-level recommendations (areas of further specification) will come from this discussion.

Other discussion items:				
Action Items:	**Name Responsible**	**Action Item**	**Date Due**	**Notes**
	Christine/Arumani	Discussion with Liaisons & determine action items.	May 2004	